1-1-1990

On the Optimal Reward Function of the Continuous Time Multiarmed Bandit Problem

José Luis Menaldi
Wayne State University, menaldi@wayne.edu

Maurice Robin
Institut National de Recherche en Informatique et en Automatique, Rocquencourt

SIAM J. CONTROL AND OPTIMIZATION
Vol. 28, No. 1, pp. 97-112, January 1990

005

ON THE OPTIMAL REWARD FUNCTION OF THE CONTINUOUS TIME MULTIARMED BANDIT PROBLEM*

JOSÉ LUIS MENALDI† AND MAURICE ROBIN‡

Abstract. The optimal reward function associated with the so-called "multiarmed bandit problem" for general Markov-Feller processes is considered. It is shown that this optimal reward function has a simple expression (product form) in terms of individual stopping problems, without any smoothness properties of the optimal reward function neither for the global problem nor for the individual stopping problems. Some results relative to a related problem with switching cost are obtained.

Key words. variational inequality, switching problem, bandit problem, dynamic programming, index policy

AMS(MOS) subject classifications. 35B37, 49A60, 49B60, 60J25, 93E20

1. Introduction. This paper deals with the properties of the optimal reward function associated with the so-called "multiarmed bandit problem." Let us recall, formally, the statement of the problem: assume that there are N independent machines. $x_i(t)$, $t \in \mathbb{R}_+$ is the state (for instance the production) of machine i. At each time t, one operates only one machine, the others being frozen. When machine i is operating, $x_i(t)$ evolves as a continuous time Markov process with a given semigroup $\Phi^i(t)$. If $i(t)$ denotes the number of the machine in operation at time t, we want to maximize a global payoff

$$J = E \int_0^\infty e^{-\alpha t} f(i(t), x_{i(t)}(t))\, dt \tag{1.1}$$

where f is a given instantaneous reward.

The multiarmed bandit problem has been studied by Gittins [4] and Whittle [8] in the discrete time case, and more recently by Varaiya, Walrand, and Buyukkoc [7] in a more general setting. Karatzas [5] studied the continuous time case when $x_i(t)$ is a one-dimensional diffusion process. The most general study is done in Mandelbaum [13], [14] who formulated the problem as the control of a multiparameter process. This approach allows, in particular, a strong formulation of the optimal process when $x_i(t)$ is a diffusion process.

In Whittle [9] it is shown that the optimal reward function has a simple expression in terms of an individual stopping problem each involving only one machine. Such an expression is shown to hold true for the diffusion bandit problem in Karatzas [5] thanks to the smoothness of the reward function which allows explicit computations.

In this paper, the main objective is to obtain such an expression when the $x_i(t)$ are general Feller processes, without smoothness properties of the optimal reward function neither for the global problem nor for the individual stopping problem.

Let us describe briefly what expression we are looking for.

* Received by the editors October 19, 1987; accepted for publication (in revised form) March 9, 1989.

† Wayne State University, Department of Mathematics, Detroit, Michigan 48202. This research was partially supported by National Science Foundation grant DMS-8601998 and Air Force Office of Scientific Research contract F49620-86-C-0111.

‡ Institut National de Recherche en Informatique et en Automatique, Rocquencourt, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay, France.

Following Whittle [9], we will use the variant of the problem where one can decide, at any time, to stop the control problem, with a reward M if this "retirement option" is chosen.

Assume that $x_i(t)$ is for each i, a Markov process with values in some space E_i, with semigroup $\Phi^i(t)$.

If $\underline{x}$ denotes the initial state of the whole set of machines, and if $u(\underline{x}, M)$ is the corresponding optimal reward function, then by applying, formally, the dynamic programming arguments, $u(\underline{x}, M)$ is shown to be the minimum solution of the following inequalities:

$$u(\underline{x}, M) \geqq e^{-\alpha t}\Phi^i(t)u(\underline{x}, M) + \int_0^t e^{-\alpha s}\Phi^i(s)f_i(x_i)\,ds \tag{1.2}$$

$$u(\underline{x}, M) \geqq M.$$

The individual stopping problems have optimal cost functions $(\phi_i(x_i, M), i = 1, N)$, where ϕ_i is the minimum solution of

$$\phi_i(x_i, M) \geqq e^{-\alpha t}\Phi^i(t)\phi_i(x_i, M) + \int_0^t e^{-\alpha s}\Phi^i(s)f_i(x_i)\,ds \tag{1.3}$$

$$\phi_i(x_i, M) \geqq M$$

when $\alpha k \leqq f_i(x_i) \leqq \alpha K, \quad \forall i, \quad \forall x_i$.

The objective is to show that

$$u(\underline{x}, M) = K - \int_M^K \sum_{i=1}^N \frac{\partial \phi_i}{\partial m}\,dm. \tag{1.4}$$

It would be nice to obtain such a formula by analytic methods, as it can be shown that (1.2) and (1.3) have a minimal solution (cf. [1], [2], [3]). However, without smoothness on ϕ_i, we do not know how to show the result by analytic methods.

Here we will use an intermediary control problem (§ 2.1) which is suitable for our objective, although it does not contain a general statement of the multiarmed bandit itself when there is no switching cost.

Using this particular interpretation of the minimal solution of (1.2), we will show (1.4) using an extension to the continuous time case of the Tsitsiklis' lemma [6]. In § 3, we investigate the problem with switching cost, showing a similar lemma; it does not seem possible, however, to obtain an expression of the optimal reward in terms of some individual problems.

2. Problem without switching cost. We start with a control problem which will provide a stochastic interpretation of (1.2).

2.1. An intermediary control problem. Let E_i, $i = 1 \cdots N$ be a family of compact metric spaces endowed with their Borel σ-algebra.

Define $E = E_1 \times \cdots \times E_N$. Throughout the paper,

$\underline{x}$ will denote an element of E, i.e.,

$$\underline{x} = (x_1, \cdots, x_N), \qquad x_i \in E_i.$$

We are given a family of Markov semigroups $\Phi^i(t)$ $i = 1, \cdots, N$, $\Phi^i(t)$ being defined and *continuous* on $C(E_i)$, the Banach space of continuous functions on E_i.[1]

[1] So, Φ^i is a Feller semigroup on $C(E_i)$, cf. Dynkin [10].

If $\Omega_i = D(\mathbb{R}_+, E_i)$, the space of right continuous, left limited functions on $\mathbb{R}_+$ with values in E_i, we denote by $Q^i_{x_i}$ the probability measure on Ω_i corresponding to Φ^1, and we define

$$\Omega = \Omega_1 \times \cdots \times \Omega_N$$

and $\{F_t\}$ the associated canonical σ-algebra.

In order to define the controlled process, we first consider the probability measure corresponding to constant trajectories for the components $j \neq i$ (i being the number of the process which is active, the others being frozen), and which gives the markovian evolution corresponding to $\Phi^i(t)$ for the component i: in other words we define

$$P_{i,\underline{x}} = \delta_{x_1} \times \cdots \times \delta_{x_{i-1}} \times Q^i_{x_i} \times \delta_{x_{i+1}} \cdots \times \delta_{x_N}. \tag{2.1}$$

Notice that, if

$$\underline{x}_t(\omega) = \omega(t) \quad \text{for } \omega \in \Omega,$$

then

$$E_{i,\underline{x}} g(\underline{x}_t) = E^i_{x_i} g(x_1, \cdots, x_{i-1}, x_i(t), x_{i+1}, \cdots, x_N)$$

where $E_{i,\underline{x}}$ (respectively, $E^i_{x_i}$) denotes the expectation with respect to $P_{i,\underline{x}}$, (respectively, $Q^i_{x_i}$).

Assume now that

(2.2) $f_i(x_i)$ is a positive function $f_i \in C(E_i)$, $\forall i$ $\alpha > 0$ a discount factor

(2.3) V will be the set of admissible controls and $v \in V \Leftrightarrow v = (\theta_n, \xi_n)_{n \geqq 0}$, $\theta_0 = 0$, where (θ_n) is an increasing sequence of F_t stopping times, ξ_n a F_{θ_n}-measurable random variable with values in $\{1, \cdots, N\}$ and we assume

(2.4) $\theta_n(\omega) \uparrow +\infty \quad \forall \omega$.

For any $v \in V$, $\underline{x} \in E$, we define, as in [11], the following sequence of probability measures on (Ω, F_∞), if $\xi_0 = i$

$$P^0 = P_{i,\underline{x}}$$

P^1 is the (unique) probability measure on (Ω, F_∞) such that

$$P^1 = P^0 \text{ on } F_{\theta_1}$$

$$P^1(\eta_{\theta_1} B \,|\, F_{\theta_1}) = P_{\xi_1, \underline{x}_{\theta_1}}(B), \quad P^0 \text{ a.s.},$$

$\forall B$ Borel subset of Ω, η_t being the shift operator,

and so on $\cdots$

P^n is similarly defined from P^{n-1}

$$P^n = P^{n-1} \quad \text{on } F_{\theta_n}$$

$$P^n(\eta_{\theta_1} B \,|\, F_{\theta_n}) = P_{\xi_n, \underline{x}_{\theta_n}}(B), \quad P^{n-1} \text{ a.s.}$$

Defining

$$\xi(t) = \xi_n \quad \text{for } t \in [\theta_n, \theta_{n+1}[, \qquad n \geqq 0.$$

We consider the discounted reward

$$J_{\underline{x}}(v) = \lim_{n \uparrow \infty} E^n_{\underline{x}} \int_0^{\theta_{n+1}} e^{-\alpha t} f(\xi(t), \underline{x}_t)\, dt$$

where

$$f(\xi, \underline{x}) = f_i(x_i) \quad \text{iff } \xi = i.$$

Actually, with the assumptions $\theta_n \uparrow +\infty$, one can know that there exists a unique probability measure $P^v_{i,\underline{x}}$ on (Ω, F_∞) such that

$$P^v_{i,\underline{x}} = P^n \quad \text{on } F_{\theta_n} \tag{2.5}$$

and one can also define our total reward by

$$J_{\underline{x}}(v) = E^v_{\underline{x}} \int_0^\infty e^{-\alpha t} f(\xi_t, \underline{x}_t)\, dt. \tag{2.6}$$

We now add another control possibility, namely the "retirement option."

Let T be the set of F_t stopping times, for $v \in V$, $\tau \in T$, and $(i, \underline{x}) \in U \times E$, we define the total reward as

$$J^M_{\underline{x}}(v, \tau) = E^v_{\underline{x}} \left\{ \int_0^\tau e^{-\alpha t} f(\xi_t, \underline{x}_t)\, dt + e^{-\alpha \tau} M \right\} \tag{2.7}$$

where M is a given constant.

We will use, as in Whittle [9], the additional assumption

$$\alpha k \leqq f_j \leqq \alpha K, \quad \forall j \in U, \tag{2.8}$$

where $k < K$ are given nonnegative constants.

The optimal reward function is

$$u(\underline{x}, M) = \operatorname{Sup} (J^M_{\underline{x}}(v, \tau), (v, \tau) \in V \times T). \tag{2.9}$$

Using a *formal* dynamic programming argument, it is easy to check that $u(\underline{x}, M)$ should solve the following inequalities

$$w(\underline{x}, M) \geqq e^{-\alpha t} \Phi^i(t) w + \int_0^t e^{-\alpha s} \Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0 \quad \forall i \in U,$$

$$w(\underline{x}, M) \geqq M, \tag{2.10}$$

$$w(\cdot, M) \text{ is a bounded measurable function.}$$

In the following section, we will show that u is actually the *minimum* solution of these inequalities (for fixed M).

Let us recall the following result (cf. Bensoussan and Robin [3], Bensoussan [1]):

THEOREM 2.1. *Under the assumption* (2.2) *there exists a minimum solution* $\bar{u}$ *of* (2.10) *in the space of bounded measurable functions. Moreover* $\bar{u}$ *is upper semicontinuous.*

Remark 2.1. In Bensoussan and Robin [3], another kind of interpretation was given for $\bar{u}(\underline{x}, M)$. The present one will be more suitable for the problem we consider.

2.2. Characterization of the optimal reward (2.9). In order to characterize $u(\underline{x}, M)$ as defined in (2.9), we introduce another switching problem, with a switching cost ε. Namely, we consider the same problem as in § 2.1, but now, at each switching time a cost ε (i.e., a reward $-\varepsilon$) is involved. This is in fact a classical switching problem (which can be considered as an impulse control problem where the state is $(\xi_t, \underline{x}_t)$, cf. Bensoussan [1], Bensoussan and Lions [2] for the general theory).

In this context, let

$$V_0 = \{v = (\theta_n, \xi_n)_{n \geqq 1}\}$$

be the set of admissible controls, θ_n, ξ_n being defined as previously.

For $(i, \underline{x}) \in U \times E$, define the reward

$$J_{i,\underline{x}}^{M,\varepsilon}(v, \tau) = E_{i,\underline{x}}^{v}\left\{\int_0^{\tau} e^{-\alpha s} f(\xi(t), \underline{x}_t)\, dt - \varepsilon \sum_{j \geqq 1} e^{-\alpha \theta_j} \chi_{\theta_j < \tau} + e^{-\alpha \tau} M\right\} \tag{2.11}$$

where $E_{i,\underline{x}}^{v}$ is defined as in § 2.1, $(v, \tau) \in V_0 \times T$, and $\chi_B(\omega)$ is the characteristic function of the set B and $\xi_0 = i$ for the construction of $P_{i,\underline{x}}^{n}$.

We also define

$$u_i^{\varepsilon}(\underline{x}, M) = \sup\, (J_{i,\underline{x}}^{M,\varepsilon}(v, \tau), (v, \tau) \in V_0 \times T). \tag{2.12}$$

Let $u^{\varepsilon} = (u_1^{\varepsilon}, \cdots, u_N^{\varepsilon})$.

From impulse control theory (cf. Bensoussan and Lions [2], [11]) we know that, for fixed M, u^{ε} is the minimum element of the set of bounded measurable functions w satisfying

$$\begin{aligned} &w_i(\underline{x}) \geqq e^{-\alpha t} \Phi^i(t) w_i + \int_0^t e^{-\alpha s} \Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0, \\ &w_i(\underline{x}) \geqq -\varepsilon + \max_j w_j(\underline{x}), \qquad (2.13) \\ &w_i(\underline{x}) \geqq M. \end{aligned}$$

Moreover, $u_i^{\varepsilon}(\underline{x}) \in C(E)$, $\forall i = 1, \cdots, N$.

We first establish the following result.

THEOREM 2.2. *Let $\underline{u}(\underline{x}, M)$ be the minimum solution of the inequalities* (2.10), *then*

$$\lim_{\varepsilon \downarrow 0} u_i^{\varepsilon}(\underline{x}, M) = \underline{u}(\underline{x}, M) \tag{2.14}$$

pointwise in $\underline{x}$.

Proof. It is clear that $u_i^{\varepsilon}(\underline{x}, M)$ increases when ε decreases, and that $u_i^{\varepsilon}(\underline{x}, M)$ is bounded (say by $(1/\alpha)\|f\| + M$). Let us define

$$\underline{w}_i = \lim_{\varepsilon \downarrow 0} u_i^{\varepsilon}.$$

From (2.13), we have

$$\underline{w}_i \geqq \max_j w_j(\underline{x}, M), \quad \forall i. \tag{2.15}$$

Hence

$$\underline{w}_i(\underline{x}, M) = \underline{w}(\underline{x}, M) \quad \forall i.$$

But $\underline{u}(\underline{x}, M)$, the minimum solution of (2.10), satisfies obviously (2.13) and therefore

$$u_i^{\varepsilon}(\underline{x}, M) \leqq \underline{u}(\underline{x}, M) \quad \forall \varepsilon, i.$$

So we deduce, when $\varepsilon \to 0$,

$$\underline{w}(\underline{x}, M) \leqq \underline{u}(\underline{x}, M). \tag{2.16}$$

But we see that $\underline{w}(\underline{x}, M)$ will also satisfy (2.10), since this is identical to (2.13) when $\varepsilon = 0$ for a function which does not depend explicitly on i.

Therefore

$$\underline{w}(\underline{x}, M) \geqq \underline{u}(\underline{x}, M).$$

Hence

$$\underline{w}(\underline{x}, M) = \underline{u}(\underline{x}, M)$$

and the theorem is proved. □

Let us define

$$u(\underline{x}, M) = \sup (J_{\underline{x}}^{M}(v, \tau), (v, \tau) \in V \times T). \tag{2.17}$$

Then we have the following Theorem.

THEOREM 2.3.

$$\underline{u}(\underline{x}, M) = u(\underline{x}, M). \tag{2.18}$$

Proof. Since $\varepsilon > 0$, we have

$$u_i^{\varepsilon}(\underline{x}, M) = \sup (J_{i,\underline{x}}^{M,\varepsilon}(v, \tau), (v, \tau) \in V_0 \times T)$$

and

$$J_{i,\underline{x}}^{M,\varepsilon}(v, \tau) \leqq J_{\underline{x}}^{M}(\tilde{v}, \tau)$$

where $\tilde{v} = ((i, 0), v)$, and $(\tilde{v}, \tau) \in V \times T$.

Therefore

$$u_i^{\varepsilon}(\underline{x}, M) \leqq u(\underline{x}, M),$$

hence

$$\underline{u}(\underline{x}, M) \leqq u(\underline{x}, M). \tag{2.19}$$

Now, for any solution w of the inequalities (2.10), one can show as in [11, Thm. VII, § 3.1] or [2b, § 6.4], that

$$w(\underline{x}) \geqq E_x^m \left\{ e^{-\alpha\theta_{m+1}\wedge\tau} w(\underline{x}_{\theta_{m+1}\wedge\tau}) + \int_0^{\theta_{m+1}\wedge\tau} e^{-\alpha t} f(\underline{x}_t, v_t)\, dt \right\}$$

for any admissible control (v, τ), $v = (\theta_i, \xi_i)_{i\geqq 0}$ where E_x^m is the expectation corresponding to the measure P_x^m associate to (v, τ) as in § 1, with $\theta_m \wedge \tau$ instead of θ_m.

From this inequality, we deduce, when $m \to +\infty$, since $w(\underline{x}) \geqq M$ and $\theta_m \wedge \tau \uparrow \tau$, that

$$w(\underline{x}) \geqq J_{\underline{x}}^{M}(v, \tau)$$

and therefore

$$w(\underline{x}) \geqq u(\underline{x}, M).$$

Finally, this gives

$$\underline{u}(\underline{x}, M) \geqq u(\underline{x}, M),$$

which, with (2.19), proves the result. □

2.3. Reduction to write off policies. Following Whittle, a write off policy is defined as a policy such that there exists a family of "write-off" sets $S_i \subset E_i$ with the following properties.

- as soon as x_i (the state of the process i) belongs to S_i, the process i is abandoned;
- one retires as soon as all the processes have been abandoned, and only then;
- before retiring, one works only with those processes which have not been abandoned.

In this section we are going to show a lemma similar to the one obtained by Tsitsiklis [6] for discrete time, showing that we can restrict ourselves to write off policies with write off sets defined by optimal stopping problems for the individual processes.

The individual stopping problems. Let us consider the optimal stopping reward

$$\phi_i(x_i, M) = \sup_{\tau} I^M_{x_i}(\tau) \tag{2.20}$$

$$I^M_{x_i}(\tau) = E^i_{x_i}\left[\int_0^\tau e^{-\alpha t} f_i(x_{it})\, dt + e^{-\alpha\tau} M\right]. \tag{2.21}$$

It is known from standard theory (see Bensoussan [1]) that $\phi_i(x_i, M)$ is the minimum element of the set of functions $w(x_i)$ satisfying

$$\begin{aligned} &w(x_i) \geqq e^{-\alpha t}\Phi^i(t) w + \int_0^t e^{-\alpha s}\Phi^i(s) f_i(x_i)\, ds, \quad \forall t > 0 \\ &w(x_i) \geqq M, \qquad w \in C(E_i). \end{aligned} \tag{2.22}$$

Let us show the following results which extend the discrete time case (cf. Whittle [9]) and the diffusion case (Karatzas [5]).

LEMMA 2.1. *Under the assumptions* (2.2)-(2.8), $\phi(x, M) = \phi_i(x_i, M)$ *has the following properties*:

(i) $\phi(x, M) = M \ \forall M \geqq K$;

(ii) $\phi(x, M) = \int_0^\infty e^{-\alpha t}\Phi^i(t) f_i(x)\, dt, \ \forall M \leqq k$;

(iii) $\forall x \in E_i$, $\phi(x, \cdot)$ *is an increasing convex function*;

(iv) $\phi(x, \cdot)$ *is Lipschitz continuous and in every* M *where the derivative exists*

$$0 \leqq \frac{\partial \phi}{\partial M} \leqq 1;$$

(v) *in every point where the derivative exists*

$$\frac{\partial \phi}{\partial M}(x, M) = E_x\, e^{-\alpha\hat{\tau}}$$

where $\hat{\tau}$ *is optimal for* (2.20), *namely*

$$\hat{\tau} = \inf(t \geqq 0, \phi(x, M) = M).$$

Proof. (i) This shows that $\tau = 0$ is optimal in (2.20) whenever $M \geqq K$. Since $f_i \leqq \alpha K$

$$\begin{aligned} J^M_x(\tau) &\leqq E_x\{(1 - e^{-\alpha\tau})K + e^{-\alpha\tau} M\} \\ &= K + E_x\, e^{-\alpha\tau}(M - K), \end{aligned}$$

clearly if $M \geqq K$, $\tau = 0$ gives the maximum value.

(ii) $f_i \geqq \alpha k$ implies

$$w_o(x) = \int_0^\infty e^{-\alpha t}\Phi^i(t) f_i\, dt \geqq k$$

and since $w_0(x) = e^{-\alpha t}\Phi^i(t) w_0 + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds$, we see that w_0 satisfies (2.22) for $M = k$.

Moreover $t = +\infty$ in (2.22) gives

$$w(x) \geqq w_0(x) \quad \forall w \text{ solution of (2.22)}.$$

(iii) If $0 \leqq \lambda \leqq 1$, then we check that

$$w_\lambda = \lambda\phi(x, m_1) + (1 - \lambda)\phi(x, m_2)$$

satisfies (2.22) for $m = \lambda m_1 + (1-\lambda)m_2$ and therefore

$$w_\lambda \geqq \phi(x, \lambda m_1 + (1-\lambda)m_2).$$

The increasing property is obvious from (2.20).

(iv) From (2.21) one has, for an arbitrary τ

$$I_x^{M+\delta}(\tau) - I_x^M(\tau) = E_x e^{-\alpha\tau}\delta$$

therefore, for $\delta > 0$

$$I_x^{M+\delta}(\tau) \leqq I_x^M(\tau) + \delta \leqq \phi(x, M) + \delta$$

implying

$$\phi(x, M+\delta) \leqq \phi(x, M) + \delta$$

and since $\phi(x, M+\delta) \geqq \phi(x, M)$, we see that

$$0 \leqq \frac{\partial^+\phi}{\partial M}(x, M) \leqq 1.$$

(v) Let $\hat{\tau} = \inf(t \geqq 0, \phi(x, M) = M)$, we know (cf. [1]) that

$$\phi(x, M) = I_x^M(\hat{\tau}).$$

Therefore, if $\delta > 0$,

$$\begin{aligned} I_x^{M+\delta}(\hat{\tau}) &= I_x^M(\hat{\tau}) + \delta E_x e^{-\alpha\hat{\tau}} \\ &= \phi(x, M) + \delta E_x e^{-\alpha\hat{\tau}} \end{aligned}$$

hence

$$\phi(x, M+\delta) - \phi(x, M) \geqq \delta E_x e^{-\alpha\hat{\tau}}$$

$$\frac{\partial^+\phi}{\partial M}(x, M) \geqq E_x e^{-\alpha\hat{\tau}}.$$

Taking $\delta < 0$, we get

$$\frac{\partial^-\phi}{\partial M}(x, M) \leqq E_x e^{-\alpha\hat{\tau}}.$$

Therefore, in M such that the derivative exists, we get the result. □

COROLLARY. *$\partial^+\phi/\partial M$ is a right continuous increasing function such that*

$$0 \leqq \frac{\partial^+\phi}{\partial M} \leqq 1,$$

$$\frac{\partial^+\phi}{\partial M}(x, M) = 1 \quad \forall M \geqq K$$

$$\frac{\partial^+\phi}{\partial M}(x, M) = 0 \quad \forall M < k.$$

Let us now define, for fixed i

$$y_i = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N)$$

$U_i(y_i, M)$ the optimal reward function when only the processes different from i are available.

From the previous section, $U_i(y_i, M)$ is the minimum solution of

$$W_i(y_i, M) \geqq e^{-\alpha t}\Phi^j(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^j(s) f_j(x_j)\, ds, \quad \forall t > 0, \quad \forall j \neq i \tag{2.23}$$

$$W_i(y_i, M) \geqq M, \quad W_i(\cdot, M) \text{ bounded and measurable.}$$

We can now state the Tsitsiklis' lemma in continuous time.

LEMMA 2.2. (*Tsitsiklis' lemma in continuous time*). *One has*

$$u(\underline{x}, M) \leqq \phi_i(x_i, M) - M + U_i(y_i, M), \quad \forall i. \tag{2.24}$$

Proof. Let $W_i(\underline{x}, M)$ be the right-hand side of (2.24). We are going to show that W_i satisfies (2.10) and since u is the minimum solution, this will show the lemma. Notice that since $U_i \geqq M$ and $\phi_i \geqq M$, we have

$$W_i(\underline{x}, M) \geqq M.$$

Moreover, since U_i does not depend on x_i,

$$e^{-\alpha t}\Phi^i(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds = \left\{ e^{-\alpha t}\Phi^i(t)\phi_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i\, ds \right\} + e^{-\alpha t}[U_i - M]$$

$$= I + II.$$

We have,

$$I \leqq \phi_i \quad \text{by (2.22)}$$

and, since $U_i - M \geqq 0$,

$$II \leqq U_i - M.$$

Then, for $j \neq i$, since ϕ_i does not depend on x_j, $j \neq i$,

$$e^{-\alpha}\Phi^j(t)\, W_i + \int_0^t e^{-\alpha s}\Phi^j(s) f_j\, ds$$

$$= \left\{ e^{-\alpha t}\Phi^j(t)\, U_i + \int_0^t e^{-\alpha s}\phi^j(s) f_j\, ds \right\} + e^{-\alpha t}[\phi_i - M] = III + IV.$$

Hence, using (2.23)

$$III \leqq U_i,$$

and

$$IV \leqq \phi_i - M \text{ since } \phi_i - M \geqq 0.$$

Therefore the lemma is proved. □

COROLLARY. *Define* $S_i^M = \{x_i \in E_i, \phi_i(x_i, M) = M\}$, *then one can restrict the policies to be write off with respect to* $(S_i^M, i = 1, \cdots, N)$.

Proof. Notice that $U_i(y_i, M) \leqq u(\underline{x}, M)$. If $x_i \in S_i^M$, then (2.24) gives, with the above inequality,

$$u(\underline{x}, M) = U_i(y_i, M)$$

which means that the optimal reward is the same as the one with $N-1$ processes where the i process has been dropped. If there exists i such that, for $\underline{x} = (x_1, \cdots, x_N)$, $x_i \notin S_i^M$, then $u(\underline{x}, M) \geqq \phi_i(x_i, M) > M$ implies that it is not optimal to retire. Finally

if $\underline{x}$ is such that $x_i \in S_i^M$, $\forall i$, then $u(\underline{x}, M) = U_i$ $\forall i$ and we can use the same argument for N-1 processes to show that

$$u(\underline{x}, M) = \phi_j(x_j, M) = M \quad \forall j. \qquad \square$$

Let us denote by V_{off}^M the set of admissible write off policies corresponding to $(S_i^M,\ i = 1, \cdots, N)$. We will use the following lemma due to Whittle (cf. [9]).

LEMMA 2.3. *If (v, τ) is a write off policy, then*

$$E_{\underline{x}}^v e^{-\alpha\tau} = \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i}$$

where τ_i is the retirement times when only the process i is available.

Proof. For the proof see Whittle [9]. $\square$

In our context, this means that, if

$$\tau_i^M = \inf(t \geqq 0,\ \phi_i(x_{it}, M) = M), \tag{2.25}$$

then, for all write off policies,

$$E_x^v e^{-\alpha\tau} = \prod_i E_{x_i} e^{-\alpha\tau_i^M}. \tag{2.26}$$

We can then deduce the product formula for u.

THEOREM 2.4.

$$U(\underline{x}, M) = K - \int_M^K \sum_{i=1}^{N} \frac{\partial \phi_i}{\partial m}(x_i, m)\, dm. \tag{2.27}$$

Proof. Let (v, τ) be an admissible write off policy with respect to $(S_i^M,\ i = 1, \cdots, N)$. We have,

$$J_{\underline{x}}^{M+\delta}(v, \tau) - J_{\underline{x}}^M(v, \tau) = \delta E_{\underline{x}}^v e^{-\alpha\tau}.$$

From the previous lemma

$$J_{\underline{x}}^{M+\delta}(v, \tau) - J_{\underline{x}}^M(v, \tau) = \delta \cdot \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i^M},$$

therefore

$$u(\underline{x}, M+\delta) \geqq J_{\underline{x}}^M(v, \tau) + \delta \prod_{i=1}^{N} E_{x_i} e^{-\alpha\tau_i^M}.$$

Note that the last term is independent from (v, τ) as far as (v, τ) is a write off policy with respect to (S_i^M). Therefore, maximizing with respect to (v, τ)

$$u(\underline{x}, M+\delta) \geqq u(\underline{x}, M) + \delta \prod_i E_{x_i} e^{-\alpha\tau_i^M}$$

which implies, for $\delta > 0$,

$$\frac{\partial^+ u}{\partial M}(\underline{x}, M) \geqq \prod_i E_{x_i} e^{-\alpha\tau_i^M}$$

and for $\delta < 0$

$$\frac{\partial^- u}{\partial M}(\underline{x}, M) \leqq \prod_i E_{x_i} e^{-\alpha\tau_i^M}.$$

Therefore, at every point where the derivatives exist, we have, thanks to the Lemma 2.1,

$$\frac{\partial u}{\partial M}(\underline{x}, M) = \prod_{i=1}^{N} \frac{\partial \phi_i}{\partial M}(x_i, M).$$

Integrating from M to K, using the fact that $u(\underline{x}, K) = K$, we get (2.27). □

Remark. From Bensoussan and Robin [3], we can show that the optimal reward of the discrete time problem converges to the $u(\underline{x}, M)$ when the time step h goes to zero. However, we have not been able to show the product formula in continuous time by letting h go to zero on the product formula of the discrete time case.

2.4. The forward induction lemma. Let us consider the discrete time version of the stopping problems (2.22). Namely, for $h > 0$, we define (dropping the index i)

$$r_h(x) = E_x \int_0^h e^{-\alpha s} f(x_s)\, ds$$

$$Q_h z = \phi(h) z$$

$$\beta = e^{-\alpha h}.$$

Then the optimal reward for the discrete stopping problem $\phi_h(x, m)$ is the unique solution of

$$\phi_h(x, m) = \max\,(r_h + \beta Q_h \phi_h, m).$$

Defining $V_h = \{\tau$, stopping times with values in $N_h = \{nh, n \geqq 0\}\}$, we can write

$$\phi_h(x, m) = \sup_{\tau \in V_h} E_x \left(\int_0^\tau e^{-\alpha s} f(x_s)\, ds + e^{-\alpha \tau} m \right).$$

The *index* is defined, as previously, as

$$M_h(x) = \inf\,(m > k,\ \phi_h(x, m) = m).$$

On the other hand, Whittle [9] shows that $M_h(x)$ has the following representation:

$$M_h(x) = \sup_{\tau \in V_h^*} \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x\, e^{-\alpha \tau}} \tag{2.28}$$

with $V_h^* = \{\tau$ stopping times with values in $N_h^* = N_h - \{0\}\}$.

The extension of the formula (2.28) to diffusion processes was done by Karatzas [5] using explicit calculation for one-dimensional processes. We are going to show the same formula in our context; the idea being to approximate the stopping problem (2.22) by a discrete time problem (like in Bensoussan–Robin [3]).

LEMMA 2.4. *Let $\phi(x, m)$ be defined as in* (2.20) (*where we drop the index i*), *and define*

$$M(x) = \inf\,(m > k,\ \phi(x, m) = m), \text{ and } V^* = \bigcup_h V_h^*,$$

then

$$M(x) = \sup_{\tau \in V^*} \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x\, e^{-\alpha \tau}}$$

where $V^* = \{\bigcup_h V_h^*\}$.

Proof. Starting with $M_h(x)$ we have

$$k \leqq M_h(x) \leqq K \quad \forall x, \quad \forall h.$$

Clearly, V_h^* is increasing as h decreases to zero and therefore $M_h(x)$ is increasing when $h \downarrow 0$. For fixed x, let

$$v(x) = \lim_{h \downarrow 0} M_h(x)$$

then

$$v(x) = \sup_{\tau \in V^*} Z(\tau), \quad \text{where } Z(\tau) = \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}}.$$

Indeed, for all ε, there exists h, such that

$$v \geqq M_h > v - \varepsilon \Rightarrow \exists \delta(\varepsilon) \text{ s.t.} \quad M_h - \delta(\varepsilon) > v - \varepsilon$$

and from the definition of M_h, we can find $\tau_h(\delta(\varepsilon))$ such that, $\tau_h \in V_h^*$,

$$M_h \geqq Z(\tau_h) > M_h - \delta(\varepsilon).$$

Therefore for all ε, there exists $\tau \in V^*$ such that $v \geqq Z(\tau) > v - \varepsilon$ proving that $v = \sup (Z(\tau), \tau \in V^*)$.

Let us prove that $v = M(x)$. Assume that $m \geqq v$, then

$$m \geqq v \geqq \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}} \quad \forall \tau \in V^*$$

$\Rightarrow$

$$m \geqq E_x \left(\int_0^\tau e^{-\alpha s} f(x_s) + e^{-\alpha \tau} m \right) \quad \forall \tau \in V^*$$

(and for $\tau = 0$, we have the equality). Therefore $m \geqq \phi(x, m)$.

But $\phi(x, m) \geqq m$, for all m implies $\phi(x, m) = m$ for all $m \geqq v$. Now assume that $m < v$.

Let us assume that for such m

$$\phi(x, m) = \sup_\tau E_x \int_0^\tau e^{-\alpha s} f(x_s) + e^{-\alpha \tau} m = m.$$

This would imply

$$m \geqq \frac{E_x \int_0^\tau e^{-\alpha s} f(x_s)\, ds}{1 - E_x e^{-\alpha \tau}} \quad \forall \tau \in V^*$$

which contradicts the assumption $m < v$.

Therefore

$$m < v \Rightarrow \phi(x, m) > m, \quad \text{hence } v = M(x). \qquad \square$$

Remark. As it was stressed in Katehakis-Veinott [12] we can also characterize $M_h(x)$ using the "restart in x-problem" for which the optimal reward function $v_h^x(\cdot)$ is given by

$$v_h^x(y) = \max (r_h(y) + \beta Q_h v_h^x,\ r_h(x) + \beta Q_h v_h)$$

and then (see [12]) we have

$$M_h(x) = v_h^x(x).$$

In continuous time, in order to define a similar problem, we can use the discrete time solution; namely, as in Bensoussan-Robin [3], we could show that for $h = 2^{-N}$

$$v_N^x(y) = v_h^x(y)$$

is increasing when $N \to +\infty$, (and bounded).

Then

$$v^x(y) = \lim_{N \uparrow \infty} v_N^x(y)$$

is the minimum solution of the inequalities

$$v^x(y) \geqq e^{-\alpha t}\phi(t)v^x + \int_0^t e^{-\alpha s}\Phi(s)f\,ds$$

$$v^x(y) \geqq v^x(x), \quad \forall y$$

$v^x(x)$ bounded measurable functions.

This is the continuous time version of the restart in x-problem.

3. The problem with switching cost. We now turn back to the case where there is a switching cost incurred at each time we change the active process. This was already considered in § 2.2 when we constructed the functions u_i^ε. Recall that this is a more or less standard impulse control problem where the underlying state is in fact $(z, \underline{x})$ where $z \in \{1, \cdots, N\}$ is the number of the active process. It would be interesting to know if a product formula like (1.4) holds. We do not know the answer, neither for the question of the optimality of some index rule. However, we can show that the concept of write off policy is still valid in this case and this gives some more information on the optimal policies than the mere interpretation of the dynamic programming condition. The reduction to write off policies will be a consequence of the following simple result, similar to the Tsitsiklis' lemma. Let us make precise some notations: we drop the ε in the optimal reward which is now

$$u(z, \underline{x}, M) = \sup\,(J_{z,\underline{x}}^M(v, \tau), (v, \tau) \in V_0 \times T) \tag{3.1}$$

$J_{z,\underline{x}}^M(v, \tau)$, V_0, T being defined as in (2.11), with $z \in \{1, \cdots, N\}$. We know that u is the minimum element of the set of bounded and measurable functions $w(z, \underline{x})$ satisfying

$$w(z, \underline{x}) \geqq e^{-\alpha t}\Phi^z(t)w + \int_0^t e^{-\alpha s}\Phi^z(s)f_z(x_z)\,ds$$

$$w(z, \underline{x}) \geqq -\varepsilon + \max_j w(j, \underline{x}), \tag{3.2}$$

$$w(z, \underline{x}) \geqq M, \quad \forall z \in \{1, \cdots, N\}.$$

We denote by

$$y_i = (x_j, j \neq i),$$

$U(z, y_i, M)$ the optimal reward when only the processes different from i are available, and when the initial active process is the process number z.

LEMMA 3.1. *We have for arbitrary* $i \in \{1, \cdots, N\}$,

$$u(j, \underline{x}, M) \leqq [\phi_i(x_i, M) - (M + \varepsilon)]^+ + U(j, y_i, M) \quad \forall j \neq i \tag{3.3}$$

$$u(i, \underline{x}, M) \leqq \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right]. \tag{3.4}$$

Proof. Let us define, for fixed i:

$$w(z, \underline{x}) = \begin{cases} [\phi_i(x_i, M) - (M+\varepsilon)]^+ + U(z, y_i, M) & \text{for } z \neq i \\ \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right] & \text{for } z = i. \end{cases}$$

We are going to show that $w(z, \underline{x})$ satisfies (3.2) and since $u(z, \underline{x}, M)$ is the minimum solution, this will prove the lemma. We have,

$$\begin{aligned} & e^{-\alpha t}\Phi^z(t) w(z, \underline{x}) + \int_0^t e^{-\alpha s}\Phi^z(s) f_z(x_z)\, ds \\ &= \left\{ e^{-\alpha t}\Phi^z(t) U(z, y_i, M) + \int_0^t e^{-\alpha s}\Phi^z(s) f_z(x_z)\, ds \right\} \\ &\quad + e^{-\alpha t}[\phi_i(x_i, M) - (M+\varepsilon)]^+ \quad \text{if } z \neq i \\ &= \left\{ e^{-\alpha t}\Phi^i(t)\phi_i + \int_0^t e^{-\alpha s}\Phi^i(s) f_i(x_i)\, ds \right\} \\ &\quad + e^{-\alpha t}\left[\max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right] - M\right] \quad \text{if } z = i. \end{aligned}$$

In the first case, thanks to (3.2), the right-hand side is less than

$$U(z, y_i, M) + e^{-\alpha t}[\phi_i(x_i, M) - (M+\varepsilon)]^+$$

i.e., less than

$$U(z, y_i, M) + [\phi_i(x_i, M) - (M+\varepsilon)]^+ = w(z, \underline{x}).$$

In the second one, thanks to (2.22) for ϕ_i, the right-hand side is less than

$$\phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right] = w(z, \underline{x}) \quad \text{if } z = i.$$

Therefore the first inequality of (3.2) is satisfied. It is obvious that $w(z, \underline{x}) \geqq M$. Now, for the second inequality of (3.2), we must check that

$$w(i, \underline{x}) \geqq -\varepsilon + \max_{j \neq i} ([\phi_i(x_i, M) - (M+\varepsilon)]^+ + U(j, y_i, M)) \tag{3.5}$$

and, for $z \neq i$

$$w(z, \underline{x}) \geqq -\varepsilon + \max \begin{cases} [\phi_i(x_i, M) - (M+\varepsilon)]^+ + U(j, y_i, M) & \forall j \neq i \\ \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right]. \end{cases} \tag{3.6}$$

But, since $\phi_i(x_i, M) - M \geqq [\phi_i(x_i, M) - (M+\varepsilon)]^+$, (3.5) is obvious from the definition of $w(i, \underline{x})$. For (3.6), since

$$U(z, y_i, M) \geqq -\varepsilon + \max_{j \neq i} U(j, y_i, M)$$

we have

$$w(z, \underline{x}) \geqq -\varepsilon + [\phi_i(x_i, M) - (M+\varepsilon)]^+ + \max_{j \neq i} U(j, y_i, M)$$

and since $[\phi_i(x_i, M)-(M+\varepsilon)]^+ \geqq \phi_i(x_i, M)-M-\varepsilon$, we also have

$$w(z, \underline{x}) \geqq -\varepsilon + \phi_i(x_i, M) - M + \max\left[M, -\varepsilon + \max_{j \neq i} U(j, y_i, M)\right].$$

Therefore, the lemma is obtained. □

Let us define the following write off sets:

$$S_i^M = \{(z, \underline{x}) \in \{1, \cdots, N\} \times E_i \text{ such that either } z = i \text{ and } \phi_i(x_i, M) = M, \text{ or } z \neq i \text{ and } \phi_i(x_i, M) \leqq M + \varepsilon\}. \tag{3.7}$$

THEOREM 3.1. *We can restrict the admissible policies to be write off with respect to* $(S_i^M, i = 1, \cdots, N)$, *in other words*

(i) *if* $\exists i$ *s.t.* $(z, \underline{x}) \notin S_i^M$, *we continue* (*i.e., we do not use the retirement option*)

(ii) *if* $\forall i$, $(z, \underline{x}) \in S_i^M$, *we retire*

(iii) *if* $(z, \underline{x}) \in S_i^M$, *the process i is abandoned.*

Proof. (i) Assume that $\exists i$ s.t. $(z, \underline{x}) \notin S_i^M$

then -either $z = i$ and $\phi_i(x_i, M) > M$ hence $u(i, \underline{x}, M) \geqq \phi_i(x_i, M) > M$,

therefore we do not retire;

-or $z \neq i$ and $\phi_i(x_i, M) > M + \varepsilon$

hence $u(z, \underline{x}, M) \geqq -\varepsilon + \max_j u(j, \underline{x}, M) \geqq -\varepsilon + \phi_i > M$.

Therefore we do not retire.

(ii) Assume that

$$\forall i, \quad (z, \underline{x}) \in S_i^M \tag{3.8}$$

and to fix the idea, take $z = N$, then (3.3) implies, since $(z, \underline{x}) \in S_1^M$, and $U(z, y_1, M) \leqq u(z, \underline{x}, M)$,

$$u(z, \underline{x}, M) = U(z, y_1, M).$$

Denote U by $U^{N-1}(z, y_1^{N-1}, M)$ to make explicit that U is the optimal reward of a problem where only the $N-1$ first components are available, i.e., $y_1^{N-1} = (x_2, \cdots, x_N)$.

Then applying again (3.3) to the $N-1$ dimensional bandit problem we get, with $i = 2$

$$u(z, \underline{x}, M) = U^{N-1}(z, y_1^{N-1}, M) = U^{N-2}(z, y_2^{N-2}, M)$$

with $y_2^{N-1} = (x_3, \cdots, x_N)$.

This process goes on until

$$u(z, \underline{x}, M) = U^1(z, x_z, M) = \phi_z(x_z, M)$$

which by the assumption and (3.4) is equal to M. Therefore we must retire if (3.8) holds.

(iii) Assume $(z, \underline{x}) \in S_i^M$

-either $z \neq i$ then (3.3) and $u(z, \underline{x}, M) \geqq U(z, y_i, M)$ implies $u(z, \underline{x}, M) = U(z, y_i, M)$ meaning that we never use again the process i

-or $z = i$ and $\phi_i(x_i, M) = M$, then (3.4) implies that either we retire, or we have

$$u(z, \underline{x}, M) = -\varepsilon + \max_{j \neq z} U(j, y_z, M)$$

meaning that we switch to another process and never use the process $z = i$. This completes the proof of Theorem 3.1. □

Acknowledgment. The authors would like to thank Professors A. Bensoussan and P. L. Chow for the useful discussions on this work.

REFERENCES

[1] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
[2a] A. BENSOUSSAN AND J. L. LIONS, *Inéquations variationnelles et problèmes d'arrêt optimal*, Dunod, Paris, 1978.
[2b] ———, *Applications des inéquations quasi variationnelles au contrôle stochastique*, Dunod, Paris, 1982.
[3] A. BENSOUSSAN AND M. ROBIN, *On the convergence of the discrete time dynamic programming equation for general semi-group*, SIAM J. Control Optim., 20 (1982), pp. 722–746.
[4] J. C. GITTINS, *Bandit processes and dynamic allocation indices*, J. Roy. Stat. Soc., 41 (1979), pp. 148–177.
[5] I. KARATZAS, *Gittins indices in the dynamic allocation problem for diffusion processes*, Ann. Probab., 12 (1984), pp. 173–192.
[6] J. N. TSITSIKLIS, *A lemma on the multiarmed bandit problem*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 576–577.
[7] P. VARAIYA, J. WALRAND, AND C. BUYUKKOC, *Extensions of the multiarmed bandit problem: the discounted case*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 426–439.
[8] P. WHITTLE, *Multiarmed bandit and the Gittins index*, J. Roy. Stat. Soc., 42 (1980), pp. 143–149.
[9] ———, *Optimization Over Time*, vol. 1, John Wiley, New York, 1982.
[10] E. B. DYNKIN, *Markov Processes*, vols. 1 and 2, Springer-Verlag, Berlin, New York, 1968.
[11] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse, Paris IX, 1978.
[12] M. N. KAHETAKIS AND A. F. VEINOTT, *The multiarmed bandit problem: decomposition and computation*, Math. Oper. Res., 12 (1987), pp. 262–268.
[13] A. MANDELBAUM, *Discrete multiarmed bandit and multi-parameter processes*, Probab. Theory Related Fields, 71 (1986), pp. 129–147.
[14] ———, *Continuous multiarmed bandits and multi-parameter processes*, Ann. Probab., 15 (1987), pp. 1527–1556.